AD NUMBER

AD136922

LIMITATION CHANGES

TO:

Approved for public release; distribution is unlimited.

FROM:

**Distribution authorized to U.S. Gov't. agencies and their contractors;
Administrative/Operational Use; JUN 1957. Other requests shall be referred to Office of Naval Research, Department of the Army, 875 North Randolph Street, Arlington, VA 22203-1995.**

AUTHORITY

ONR ltr, 28 Jul 1977

RESEARCH ON THE DEVELOPMENT OF

PERFORMANCE CRITERIA

Technical Report VIII

THE PREDICTABILITY OF RATINGS AS A FUNCTION OF INTER-RATER AGREEMENT

Prepared for

Psychological Sciences Division
Personnel and Training Branch
Office of Naval Research
Department of the Navy

by

Human Factors Research, Incorporated
Los Angeles, California

June, 1957

RESEARCH ON THE DEVELOPMENT OF

PERFORMANCE CRITERIA

Technical Report VIII

THE PREDICTABILITY OF RATINGS AS A FUNCTION OF INTER-RATER AGREEMENT

by

Donald N. Buckner

June, 1957

Project Personnel

Albert Harabedian
Robert R. Mackie
Forrest K. Strayer
Clark L. Wilson, Principal Investigator

Prepared for

Psychological Sciences Division
Personnel and Training Branch
Office of Naval Research
Department of the Navy

Contractor

Human Factors Research, Inc.
Los Angeles, California
Project: NR 153-625
Contract: Nonr 1241(00)

TABLE OF CONTENTS

SUMMARY AND CONCLUSIONS

The hypothesis tested was that high agreement among the ratings assigned the same men by different raters does not necessarily imply predictable ratings.

Ratings by three superior officers (Officers and Chief Petty Officers) of 100 submariners serving aboard 21 different submarines were divided into four samples so as to achieve four levels of inter-rater agreement (.00, .69, .84 and .94). Correlations were then computed within each sample between three predictor variables (Submarine School Class Standing and the Navy General Classification and Mechanical Aptitude Tests) and the mean of the three ratings assigned to each ratee.

The hypothesis was supported by the results. None of the six correlations between the predictor variables and the ratings for which the inter-rater agreement estimates were high (.84 and .94) was significantly different from zero. Four of the six correlations computed for the low agreement ratings (.00 and .69) were significantly different from zero, one at the .01 level and three at the .05 level.

It was concluded that high inter-rater agreement does not necessarily imply predictability and may indicate a lack of it. Low agreement, on the other hand, may in some cases indicate predictability and possibly validity.

THE PROBLEM

When ratings are used as criterion measures, more "ultimate" criteria of performance are generally not available for validating them; indeed if a more ultimate measure were available, ratings probably would not be employed in the first place. It is necessary, therefore, either simply to accept the ratings as valid or to seek some indirect indication of their validity. One such indication often employed is the reliability of the ratings as shown by the amount of agreement among scores assigned the same ratees by different raters. Another is the predictability of the ratings, or the extent to which they correlate with measures to which they should be related, according to logic or the results of previous research.

The hypothesis tested here is that inter-rater agreement and predictability may be incompatible indications of validity. More specifically, it is hypothesized that high inter-rater agreement is not necessarily indicative of predictability and that disagreement among raters, on the other hand, may be associated with predictability and possibly validity.

METHOD

A sample of 100 submariners was divided into four equal groups according to the degree to which the ratings assigned them by four superiors were in agreement; thus four different levels of inter-rater agreement were obtained. Correlational analyses were then performed within each group to determine the relative predictability of the ratings from scores on two aptitude tests and from final class standing at the Submarine School, New London.

The Rating Scale. Assessments of the qualifications of candidates for the Submarine School are routinely made by psychiatrists on the staff of the U.S. Navy Medical Research Laboratory. The rating scale employed in this study was originally designed for the purpose of gathering ratings to determine the validity of these assessments for subsequent performance aboard submarines.

Since the rating scale and its development have been described fully in other reports (4, 5, 6), only a brief description will be given here. A general trait scale was developed containing 10 traits considered to pertain to the technical aspects of a man's job and 10 considered to pertain to the personal adjustment aspects. The format of the scale, an example of which is given on the following page, was designed so that the rater assigned scores to all the men he was rating on one trait at a time. He assigned the ratings on a scale of 25 hypothetical submariners of the same rate and pay grade as the men being rated. Each rater assigned his ratings independently.

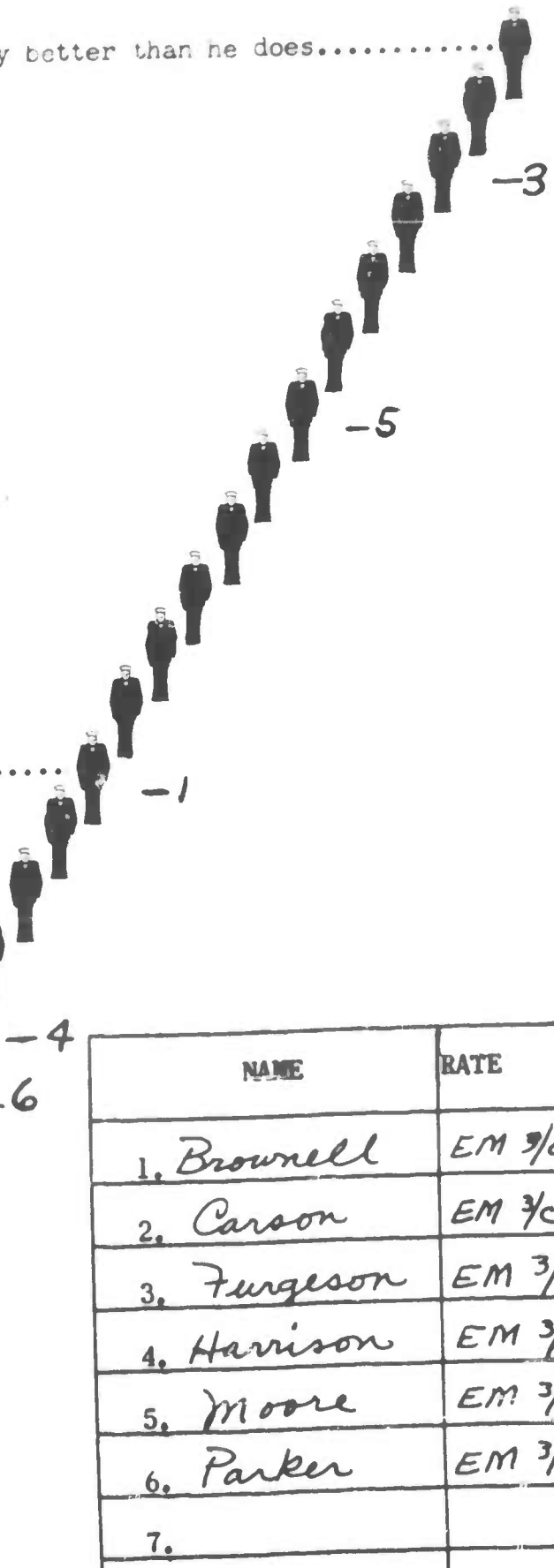
For most of the analyses reported here, only the means of the ratings assigned each ratee by each of his three raters on the ten technical competence traits were employed.

The Samples. 171 men aboard 21 different submarines of the Pacific Fleet were rated by three of their superiors, either by two Officers and a Chief Petty Officer (CPO) or one Officer and two CPOs. The raters on each boat were selected on the basis of their professed knowledge of the men to be rated.

In an attempt to control the length of time the raters had known the ratees, those men who had been aboard their boats for a period of at least ten months were selected from this total sample for the investigation reported here. Since there were 97 such men, an additional three men were randomly selected from the group that had been aboard for nine months and added to the sample to make it an even 100 ratees.

The differences between the means of the ten technical competence trait ratings assigned each ratee by the three men rating him and the mean of those three means were squared and added together, yielding what was called an "agreement" score for each ratee. That is, a ratee's agreement score indicated the extent to

ABILITY TO UNDERSTAND INSTRUCTIONS
(SAMPLE)



NAME	RATE
1. Brownell	EM 3/c
2. Carson	EM 3/c
3. Furgeson	EM 3/c
4. Harrison	EM 3/c
5. Moore	EM 3/c
6. Parker	EM 3/c
7.	
8.	

which his three raters agreed in their ratings of him. The distribution of these scores was divided at the 25th, 50th, and 75th centiles to yield four groups of 25 ratees each. Hereafter these experimental samples will be called the High Agreement (HA), Moderate Agreement (MA), Moderate Disagreement (MD), and High Disagreement (HD) groups.

Estimates of Inter-Rater Agreement. The means of the three ratings assigned to each ratee were used in the correlational analyses performed to test the hypothesis and, thus, were regarded as the ratees' "true" scores. The deviations of the three ratings assigned a man about his "true" score were then regarded as error. To estimate inter-rater agreement, these deviations were squared, summed and treated as the error variance term in the basic equation for the coefficient of reliability (2). The variance of the mean ratings or "true" scores was used as the total variance term in the equation. The inter-rater agreement estimates obtained in this manner are given in Table 1.

Table 1

ESTIMATES OF INTER-RATER AGREEMENT
FOR THE FOUR EXPERIMENTAL SAMPLES

Sample	r_{rr}
High Agreement	.94
Moderate Agreement	.84
Moderate Disagreement	.69
High Disagreement	.00

Because of the way the samples were selected, the inter-rater agreement estimates for the two agreement samples were higher than the estimates for the disagreement samples. The important thing to note is that statistics such as are shown in Table I are often presented to indicate the relative acceptability of obtained ratings as criterion measures and that the ratings showing the higher inter-rater agreement estimates would probably be preferred by most researchers conducting validation studies.

The Criteria. Three measures were used to compare the predictability of the ratings assigned to the men in the four samples: the Navy General Classification Test (GCT), the Navy Mechanical Aptitude Test (MECH), and Submarine School Class Standing (SSS). Previous research has shown that these variables are significantly related to performance aboard submarines as measured by ratings, check lists, and job sample performance tests (7). Submarine School Class Standing, which is based on a composite of written achievement test scores and instructor ratings and has an estimated reliability of .90, was found to correlate higher with scores on the shipboard criteria than any of a variety of predictor variables studied. It was selected from the measures available for this study, therefore, as

the variable most likely to be related to the ultimate criterion and as probably the best indicator of the validity as well as the predictability of the ratings.

Table 2 shows the means and standard deviations of the scores of the men in the four samples on the three predictor variables. Only the difference between the variances of the scores of the HA and HD samples on the MECH was significantly different from zero. Because of the number of tests made, one such result was expected by chance alone; therefore, the samples were considered to have been obtained from the same population with respect to these variables.

Table 2

MEANS AND STANDARD DEVIATIONS OF SCORES ON THE PREDICTOR VARIABLES FOR THE FOUR SAMPLES

		SSS		GCT		MECH	
Sample	N	M	σ	M	σ	M	σ
High Agreement	25	53.80	29.35	58.84	6.77	60.56	11.12
Moderate Agreement	25	45.52	32.20	59.62*	6.70	57.08	8.67
Moderate Disagreement	25	48.64	30.81	61.56	6.90	58.16	9.75
High Disagreement	25	52.68	27.30	58.80	6.48	59.52	6.48

* N=24

Length of Time on Board. Table 3 shows that the four samples were also homogeneous with respect to the amount of time the ratees in each group had spent aboard the submarine on which they were rated.

Table 3

MEANS AND STANDARD DEVIATIONS OF LENGTHS OF TIME SPENT ON BOARD FOR THE FOUR SAMPLES

	Months	
Sample	M	σ
High Agreement	12.64	2.33
Moderate Agreement	13.16	2.19
Moderate Disagreement	12.68	2.36
High Disagreement	13.32	2.84

Description of the Ratings Assigned. Table 4 shows the means and standard deviations of the ratings assigned the men in the four experimental samples. None of the differences between means or variances was significantly different from zero, although there was a tendency for the ratings of the HD group to be slightly less variable.

Table 4

MEANS AND STANDARD DEVIATIONS OF THE RATINGS ASSIGNED THE FOUR SAMPLES

Sample	N	M	σ
High Agreement	25	16.00	3.69
Moderate Agreement	25	13.54	4.70
Moderate Disagreement	25	14.65	4.03
High Disagreement	25	14.97	3.28

Correlational Analyses. Scores on the Navy GCT and MECH aptitude tests and Submarine School Class Standing were correlated with the mean of the ratings assigned each ratee by the three superior officers who rated him. Separate analyses were performed for each of the four experimental groups. The score actually used for Submarine School Class Standing was the proportion of men in his class the ratee exceeded; thus a man who was first in his class received a score of 1.00 while a man who was last had a score of .00. Pearson product-moment coefficients were computed.

RESULTS

Relative Predictability of the Ratings

The results of the correlational analyses are shown in Table 5.

Table 5

CORRELATIONS BETWEEN SCORES ON THE PREDICTOR VARIABLES AND RATINGS OF THE MEN IN THE FOUR EXPERIMENTAL SAMPLES

Sample	r_{rr}	SSS	GCT	MECH	N
High Agreement	.94	.05	-.23	-.07	25
Moderate Agreement	.84	.29	-.14	.02	25
Moderate Disagreement	.69	.61**	.43*	.42*	25
High Disagreement	.00	.43*	.17	.18	25

* significant at the .05 level
** significant at the .01 level

The trend was the same for all three predictor variables. As the inter-rater agreement estimates decreased from the HA to the MD sample, the correlations between the predictors and the ratings increased. Even for the sample for which the inter-rater agreement estimate was .00, the correlation between Submarine School Class Standing and the ratings was significantly different from zero (.05 level).

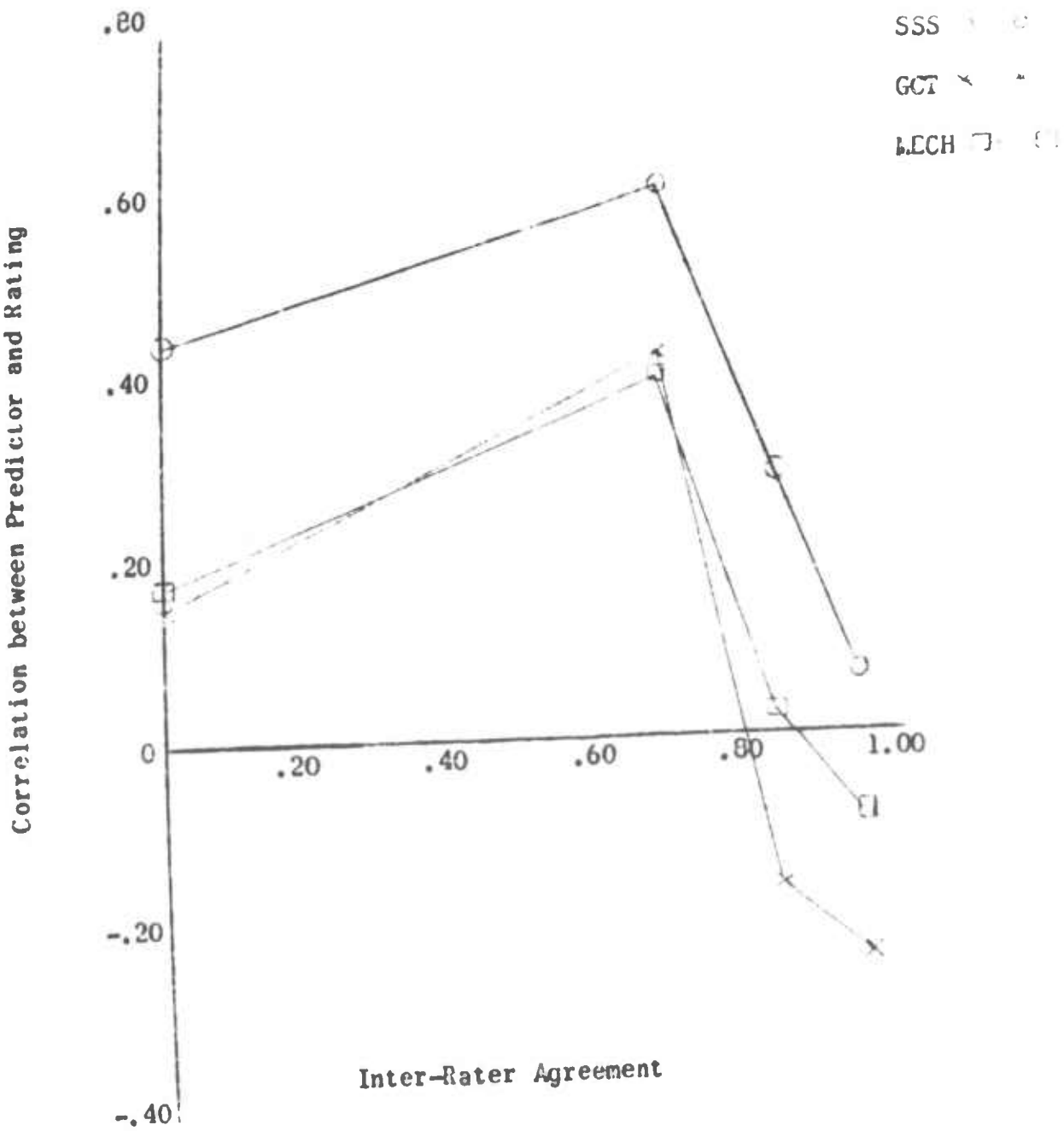
Figure 1 shows more clearly that the relationship between predictability and inter-rater agreement may be curvilinear. Predictability -- as indicated by the correlations between SSS, GCT, and MECH scores and the mean rating assigned each man -- increased as inter-rater agreement increased to about .70. At that point, the correlations dropped very rapidly to insignificant values.

It should be pointed out that there were no data points between inter-rater agreement estimates of .00 and .69. It appears conceivable from the slope of the curves between the estimates of .69 and .94 that the correlations might have been

even higher for inter-rater agreement estimates between .40 and .60.

Figure 1

CORRELATION BETWEEN SCORES ON SSS, GCT, AND MECH
AND MEAN TECHNICAL COMPETENCE RATING
AS A FUNCTION OF INTER-RATER AGREEMENT



A similar analysis was performed using the mean of the ratings on the ten personal adjustment traits. The results of this analysis are shown in Table 6.

The inter-rater agreement estimate for the MA group was slightly higher than that for the HA sample. The absolute magnitude of the sum of the squared deviations of the three ratings assigned a man about his mean rating (or "true" score) was greater for the MA than for the HA sample. However, the variance of the mean ratings of the MA group was also greater which accounts for the higher inter-rater agreement estimate.

Table 6

CORRELATIONS BETWEEN THE PREDICTOR VARIABLES AND PERSONAL ADJUSTMENT RATINGS FOR THE FOUR SAMPLES

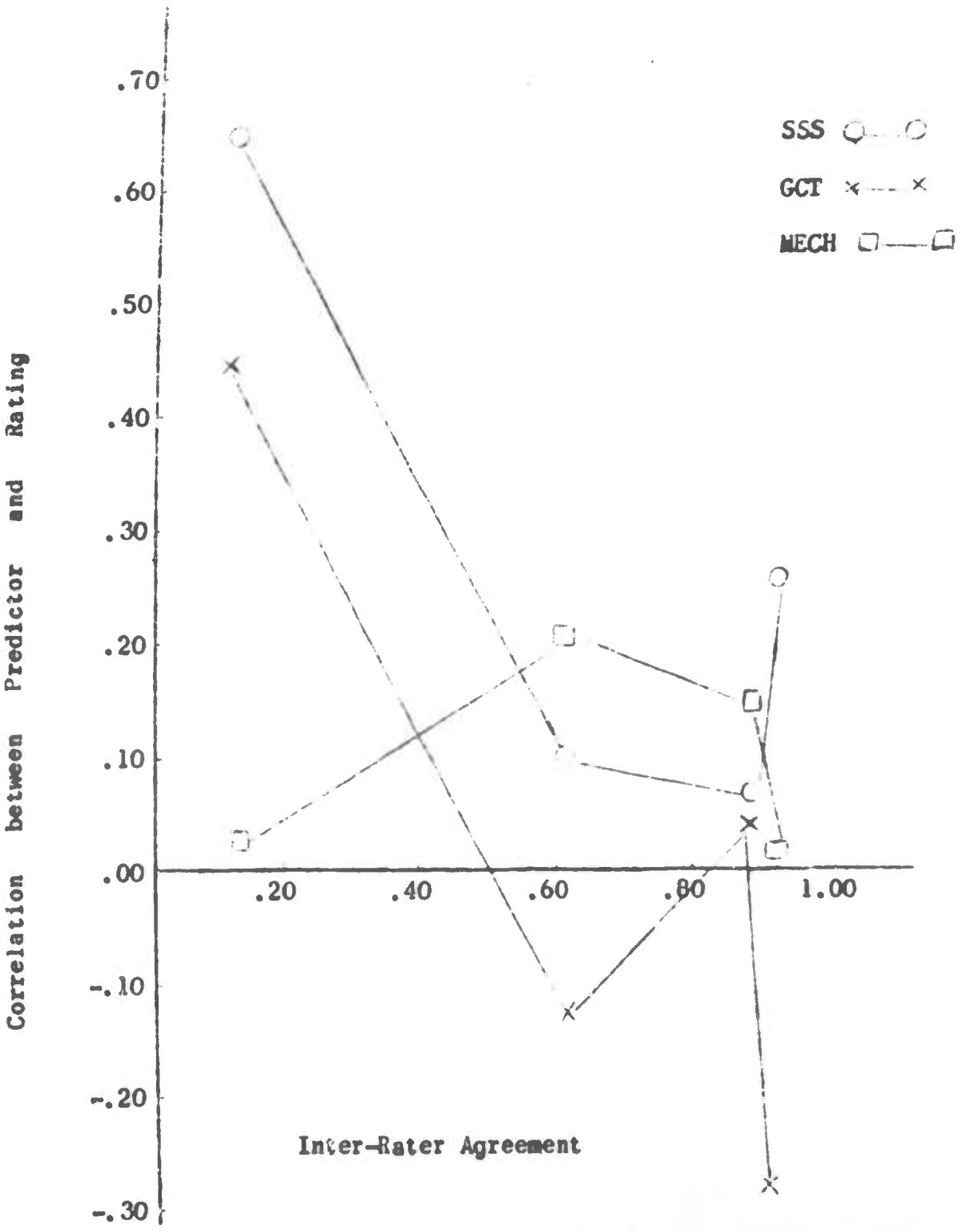
Sample	N	Err	SSS	GCT	MECH
High Agreement	25	.86	.05	.02	.16
Moderate Agreement	25	.90	.25	-.27	.01
Moderate Disagreement	25	.61	.08	-.12	.21
High Disagreement	25	.12	.65**	.44*	.06

* significant at the .05 level
** significant at the .01 level

Only two of the correlations were significantly different from zero (.05 level) and both were in the sample for which the inter-rater agreement estimate was lowest (.12). There was no apparent tendency, comparable to that found with the technical competence ratings, for predictability to increase as inter-rater agreement decreased from the two agreement groups to the MD group. Figure 2, on the following page, shows these same results in graphic form.

Figure 2

CORRELATION BETWEEN SCORES ON SSS, GCT, AND MECH
AND MEAN PERSONAL ADJUSTMENT RATING AS A FUNCTION OF INTER-RATER AGREEMENT



Investigation of Possible Sources of Predictable Variance

Additional analyses were performed in an effort to locate the source of the predictable variance in the ratings for which the inter-rater agreement estimates were low, i.e., the MD and HD groups. Only the ratings on the ten technical competence traits were used in these analyses.

First it was hypothesized that the more extreme rating with respect to the mean of the three assigned a ratee was contributing more predictable variance than the other two. The hypothesis was based on the idea that a more deviant rating might possibly indicate better observations of ratee behavior. The procedure employed in testing the hypothesis was as follows: the ratings assigned the 50 ratees in the combined MD and HD samples were plotted on a large chart. Inspection showed that all three raters disagreed in their evaluations of some men. In the case of others, two of the raters were in substantial agreement and only the third disagreed. The 25 ratees (half of the combined sample) for whom this latter pattern was most pronounced were selected for study. Scores on the predictor variables were then correlated both with the extreme rating assigned each of these 25 men and with the mean of the other two ratings given them. The results are shown in Table 7.

Table 7

CORRELATIONS BETWEEN SCORES ON THE PREDICTOR VARIABLES AND THE ONE DISAGREE RATING AND THE MEAN OF THE TWO AGREE RATINGS

Ratings	Number of Ratees	SSS	GCT	MECH
One Disagree	25	.50*	.56**	.50*
Two Agree	25	.35	.30	.41*

* significant at the .05 level
** significant at the .01 level

The means of the ratings assigned by the two raters who were in close agreement were less predictable than the ratings assigned by the rater who disagreed. (It

should be pointed out that the means and variances of these two samples of ratings were not significantly different.)

The same sort of analysis was performed using the entire combined MD and HD samples. In this case, of course, the agreement between the "two agree" raters was not as great, and in some cases the rating of the one "disagree" rater was not much farther removed from the mean of the three ratings than the rating given by one of the "two agree" raters. As shown in Table 8, the two sets of ratings were almost equally predictable from SSS. However, scores on the GCT and MECH variables correlated significantly (.05 level) with the more deviant rating and not with the mean of the ratings assigned by the two raters who were in closer agreement in their evaluations.

Table 8

CORRELATIONS BETWEEN SCORES ON THE PREDICTOR VARIABLES AND THE ONE DISAGREE AND THE MEAN OF THE TWO AGREE RATINGS (for the MD and HD samples combined)

Ratings	Number of Ratees	SSS	GCT	MECH
One Disagree	50	.43**	.30*	.29*
Two Agree	50	.44**	.19	.25

* significant at the .05 level
** significant at the .01 level

There was a slight but statistically insignificant tendency for the raters who assigned the more deviant ratings to be Chief Petty Officer rather than Officer raters. Correlational analyses indicated that the ratings assigned by these enlisted raters were also more predictable from scores on the two aptitude tests than were the ratings assigned by officers. However, as can be seen in Table 9, the two correlations with Submarine School Class standing were equal.

Table 9

CORRELATIONS BETWEEN SCORES ON THE PREDICTOR VARIABLES AND THE DISAGREE RATINGS ASSIGNED BY OFFICER AND ENLISTED CPO RATERS (for the MD and HD samples combined)

Raters	Number of Raters	SSS	GCT	
CPOs	28	.42*	.41*	.44*
Officers	22	.42*	.01	-.21

* significant at the .05 level

DISCUSSION

The results indicate that high inter-rater agreement does not necessarily imply predictability in ratings. Whether or not the results can be interpreted to mean that inter-rater agreement is not necessarily a good index of the validity of ratings depends on whether or not one is willing to accept the assumption that the predictor variables employed in the study are positively related to the ultimate criterion of the performance that was rated. Even if that assumption cannot be made, the results indicate that at least in some instances inter-rater agreement and predictability would yield incompatible indications of the validity of ratings.

Submarine School Class Standing has been shown to be more highly related to various criteria of shipboard performance than any of a variety of predictor variables studied (7). Thus, the assumption with respect to the relationships between the predictor variables used in this study and the ultimate criterion of performance aboard submarines is probably more tenable for the SSS variable. It is interesting to note in the light of this that it was also the variable that showed the most significant positive relationships with the ratings for which the inter-rater agreement estimates were low.

The explanation of the results proposed here is based first on the assumption that ratee behavior in most performance rating situations is not entirely consistent from one time to the next with respect to particular traits, primarily because no effort is made to control the physical and psychological environment during the period the ratings are designed to cover. To be valid, then, ratings must reflect these inconsistencies.

On the other hand, even if ratees behaved entirely consistently, ratings of them would not necessarily be in agreement since raters use different criteria in rating on the same trait (3). The second assumption, then, is that these criteria employed by different raters are all valid and the differences in ratings reflected

by them are also valid. This second assumption implies that part of achieving the ultimate in performance is satisfying the demands of various superiors by behaving in different ways.

On this basis, high agreement among ratings could imply a poor sampling of observations of ratee behavior by raters, a poor sampling of raters in terms of the criteria they use to evaluate particular traits, or both; whereas disagreement among the ratings assigned to the same men by different raters could indicate that a more representative sample of observations and raters was obtained. This is contrary to the usual rationale for the presence or absence of rater agreement.

It is obvious on the basis of these two assumptions that high inter-rater agreement could also indicate validity. If a ratee knew the criteria his superiors were going to employ in rating him, for example, he could -- assuming he had sufficient control of his behavior regardless of the environmental situation -- behave so as to satisfy them. It is assumed, however, that the majority of men are not entirely aware of the nature of their superiors' criteria, and even if they were, they neither have adequate control over their behavior in all situations nor are obsequious enough to attempt to satisfy all of their superiors' demands.

In summarizing the bases of unreliability in ratings, Ghiselli and Brown say, "The indication is that raters disagree primarily because they observe the individuals to be rated in different situations and under different conditions, and because they use different criteria for judging the same trait or characteristic." (Note that the two factors which they say contribute to a lack of agreement in ratings are assumed in the explanation given here to contribute to their validity, as long as they are accurately reflected in the ratings.) They continue by saying,

"It follows from this evidence that reliability of ratings can be considerably increased by having the raters observe the individuals under similar conditions,

and by providing techniques for making the ratings that will increase the likelihood that the traits or characteristics being judged will be evaluated on the same basis." (1)

While it is probably true that having raters observe ratees under similar conditions would increase inter-rater agreement, it might also serve to decrease validity for the ultimate criterion by failing to take into account all of those on-the-job situations in which individuals perform and the interaction between ratees and situations.

Certain members of a work group might react favorably in one situation and unfavorably in another. Certainly with the variety of situations individuals face from day to day regardless of their occupations, they could not be expected to react favorably in all of them. Submariners, for example, live in a threatening environment faced with an infinite number of unique situations. Officers and CPOs cannot observe their men under similar conditions simply because of the physical layout of a submarine and because of the diverse jobs individuals in the same gang perform. To develop a method whereby the ratees could be observed under similar conditions, even if it were possible, would probably imply the exclusion of critical situations in which a man's behavior would have potentially the greatest significance as far as his contribution to the effectiveness of the boat is concerned.

Increasing inter-rater agreement by having raters observe ratees under similar conditions might, therefore, defeat the more important purpose of obtaining valid ratings. The results reported here appear to support this reasoning.

It is not being suggested that high inter-rater agreement always implies a lack of predictability. Essentially none of the variance in the high agreement ratings and only a portion of the variance in the low agreement ratings was predictable from scores on the three variables used in this study. It is conceivable that both of these sources of variation could be predicted from other types of measures.

CONCLUSIONS

Since more ultimate criteria of performance are seldom available for determining the validity of performance ratings, some indirect indication of their validity is often employed. One such indication is the agreement among the ratings assigned to the same men by different raters. Another is the predictability of the ratings from measures to which they should be related, according to logic or the results of previous research.

It is concluded on the basis of the study reported here that these two indications are not necessarily compatible. Inter-rater agreement may not be a good index of predictability. Ratings for which the inter-rater agreement estimate is low may be more predictable than ratings for which that estimate is high.

Additional studies should be performed to determine whether or not these results represent a chance occurrence. If they do not, further studies would indicate the circumstances under which comparable results may be obtained.

REFERENCES

1. Ghiselli, E. E. and Brown, C. W. Personnel and Industrial Psychology. New York: McGraw-Hill, 1948, p. 90.

2. Guilford, J. P. Fundamental Statistics in Psychology and Education. 2nd ed. New York: McGraw-Hill, 1950, p. 476.

3. Guilford, J. P. Psychometric Methods. 2nd ed. New York: McGraw-Hill, 1954, p. 295.

4. Mackie, R. R. Submarine Personnel Selection, Technical Report No. 3, Several Treatments of Inter-Group Differences and Their Effect on the Reliability and Predictability of Ratings. Los Angeles: Management and Marketing Research Corporation, 1954, Nonr 1113(00).

5. Mackie, R. R., Clegg, D. A. and Buckner, D. N. Submarine Personnel Selection, Technical Report No. 2, Validation and Revision of the U. S. Medical Research Laboratory Personal History Form. Los Angeles: Management and Marketing Research Corporation, 1954, Nonr 1113(00).

6. Mackie, R. R., Strayer, F. K. and Buckner, D. N. Submarine Personnel Selection, Technical Report No. 1, Validation of the U. S. Medical Research Laboratory Personnel Assessments. Los Angeles: Management and Marketing Research Corporation, 1954, Nonr 1113(00).

7. Mackie, R. R., Wilson, C. L. and Buckner, D. N. Research on the Development of Shipboard Performance Measures, Part V: Interrelationships between Aptitude Test Scores, Performance in Submarine School, and Subsequent Performance in Submarines as Determined by Ratings and Tests. Los Angeles, Management and Marketing Research Corporation, NS onr 70001 and Nonr 1241(00), 1954.